ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Water Quality Analysis Using Multiple Linear Regression.

Nishil Shah*, Srishti Tiwari, and Vidushi Vashishtha.

School of Computing Science and Engineering, VIT University, Vellore-632014, India

ABSTRACT

Water quality discusses about characteristics of water which include physical, biological and radiological characteristics. As the industries are growing up more and more its adverse effect can be seen in water since, most toxic pollutants- human waste are discharged from industries and homes into the water system resulting in higher amount of water quality degradation. So, to overcome this issue we have presented a data mining technique called as “multiple linear regression” to check whether the water is good for drinking or not. This analysis is conducted over the dataset obtained from the water sources of Vellore district.

Keywords: Multiple linear regressions, Water Quality, Drinking Purpose, Data Mining.

**Corresponding author*

INTRODUCTION

Water Quality is one of the greatest issue of the world which is getting most exceedingly bad as the time is passing. Water quality is very much effected by the day to day activities of human being and also with industrial runoff. The most critical of the common impacts are topographical, hydrological and climatic, since these factor affect the water degradation level at most. Their impact is by and large most prominent when accessible water amounts are low and greatest use must be made of the restricted assets; for instance, high saltiness is an incessant issue in dry and waterfront territories. On the off chance that the monetary and specialized assets are accessible, seawater or saline groundwater[1] can be desalinated however as a rule this is not achievable. In this way, despite the fact that water may be accessible in satisfactory amounts, its unsatisfactory quality restrains the uses that can be made of it. Despite the fact that the common biological community is in congruity with regular water quality, any critical changes to water quality will for the most part be troublesome to the environment.

Waterfront landfills and risky waste destinations represent a potential ecological danger to surface water bodies through the trading of groundwater born contaminants. A large number of these locales are found adjoining harbours, inlets, estuaries, wetlands, and other beach front environments. Universally, prerequisite for freshwater will keep on rising altogether over the coming decades to address the issues of expanding populaces, developing economies, changing ways of life and advancing utilization designs. This will significantly enhance the weight on constrained common assets and biological communities. Risky water and sanitation represent just about one tenth of the worldwide weight of malady (Fewtrell et al., 2007). All out 768 million and 2.5 billion individuals on the planet are living without access to clean water and appropriate sanitation[2], separately (WHO, 2002; WHO and UNICEF, 2013a).

As per the World Commission on water for the 21st century, more than half of the world's significant waterways are exhausted and sullied to the degree that they undermine human wellbeing and toxin the encompassing biological systems. Debased drinking water can bring about different maladies, for example, typhoid fever, looseness of the bowels, cholera and other intestinal ailment.

METHODS

In the literature we came across three of the major techniques used for water quality analysis for one of the major river[3].

- Hierarchical Cluster Analysis
- Water quality Index
- Correlation coefficient

Hierarchical Cluster Analysis

Progressive bunch investigation (HCA) was utilized to examine the water quality information for spatial and worldly contrasts. The HCA was connected to a subgroup of the dataset to assess their handiness to order the stream water tests, and to distinguish suitability for drinking water reason. The separation bunch speaks to the level of relationship between components. The bring down the worth on the separation group, the more huge is the affiliation. HCA[4] is the most widely recognized methodology, which gives natural comparability connections between any one example and the whole information set, furthermore, is generally appeared by a dendrogram. To order the water quality in examining stations and to focus the wellspring of contamination, HCA with Ward technique, Euclidean separation in view of the institutionalized mean of the physico-synthetic parameters were utilized.

Water Quality Index

In light of the WQI an appraisal was made whether the stream water was worthy for local utilize and drinking reason. Hence, this examination is to a great degree essential. Additionally, individuals living in these territories can focus from which some portion of the waterway they can draw the best quality water. Water quality has been evaluated utilizing Water Quality Index (WQI)[5] created by the U.S. National Sanitation Foundation Water Quality Index (NSF WQI) in 1970. This record has been broadly tried on field and connected to information from various diverse land ranges everywhere throughout the world keeping in mind the end goal to ascertain Water Quality Index (WQI) of different water bodies. Basic contamination parameters were viewed as for registering WQI. Expression for NSF WQI is given by p NSF WQI = ∑ Wi Ii i=1 Where Ii is the sub-record for ith water quality parameters; Wi is the weight (in regards to noteworthiness) joined with ith water quality parameter; p is the quantity of water quality parameter.

Classification of Water Quality Index

Table 1: Water Quality Range

RANGE	QUALITY
90-100	EXCELLENT
70-90	GOOD
50-70	MEDIUM
25-50	BAD
0-25	VERY BAD

Correlation Coefficient

In the present study, relationship coefficient was utilized to distinguish the exceptionally connected water quality parameters. This can assist in with selecting the medications to minimize contaminations in waterway water There was no noteworthy relationship between water temperature and the other physical parameters with the exception of DO ($r = -0.67$; $p < 0.01$). Water temperature related adversely with the DO and absolutely with TDS and SS, with the last demonstrating a positive connection with BOD. The EC, TDS, SS and TS showed positive solid relationship with every one of the parameters aside from DO and Oil and oil. Solid positive connection of EC with BOD ($r = 0.94$; $p < 0.01$) and COD ($r = 0.88$; $p < 0.01$) bolstered the vicinity of wastewater originating from commercial ventures as the boss causative variable for oceanic contamination (Dike et al. 2013). It is clear from the outcomes that the DO was contrarily associated with every one of the variables and was not decidedly corresponded with any of the concentrated on parameter. . The DO displayed negative relationship with BOD ($r = -0.81$; $p < 0.01$), COD ($r = -0.68$; $p < 0.05$) and oil and oil ($r = -0.27$) – diminish in DO focus is connected with oxidation of re-suspended natural matter. Negative relationship in the middle of DO and NO3-N ($r = -0.67$; $p < 0.01$) was seen because of high release, which builds centralization of DO in the interstitial water as a result of expanded turbulence that diminishes the anoxic environment required for de-nitrification. The NO3-N indicated critical positive relationship with BOD ($r = 0.93$; $p < 0.01$) recommends the expansion of these supplement to River from natural waste and sewage.

EXPERIMENTAL

Multiple Linear regression

Straight relapse endeavours to demonstrate the relationship between two variables by fitting a straight comparison to watched information. One variable is thought to be a logical variable, and the other is thought to be a subordinate variable. Case in point, a modeller may need to relate the weights of individuals to their statures using an immediate backslide model. Before endeavouring to fit a straight model to watched information, a modeller ought to first figure out if or not there is a relationship between the variables of hobby. This does not as a matter of course infer that one variable causes the other (for instance, higher SAT scores[6] do not cause higher school grades), yet that there is some huge relationship between the two variables. A scatter plot can be a useful apparatus in deciding the relationship's quality between two variables. In the event that there appears to be no relationship between the proposed illustrative and subordinate variables (i.e., the scatter plot does not show any expanding or diminishing patterns), then fitting a straight relapse model to the information likely won't give a valuable model. A significant numerical measure of relationship between two variables is the correlation coefficient, which is a worth between - 1 and 1 showing the affiliation's quality of the watched information for the two variables.

A straight relapse line has a mathematical statement of the form $Y = a + bX$, where X is the illustrative variable and Y is the indigent variable. The line's slant is b, and a is the capture (the quality of y when $x=0$).

Multiple relapse is an expansion of straightforward direct relapse. It is utilized when we need to foresee[7] the estimation of a variable in view of the estimation of two or more different variables. The variable we need to foresee is known as the reliant variable (or now and again, the result, target or basis variable). The variables we are utilizing to anticipate the indigent's estimation variable are known as the free variables (or once in a while, the indicator, logical or relapse variables.

For sample, you could utilize various relapse to comprehend whether exam execution can be anticipated in view of modification time, test nervousness, address participation and sexual orientation.

Multiple relapse additionally permits you to focus the general fit (fluctuation clarified) of the model and the relative commitment of each of the indicators to the aggregate difference clarified. For instance, you might need to know the variety's amount in exam execution can be clarified by update time, test nervousness, address participation and sex "all in all", additionally the "relative obligation"[8] of every free variable in clearing up the change. Different direct relapse has the accompanying equation.

The model is direct on the grounds that it is straight in the parameters. The model portrays a plane in the three-dimensional space of. The parameter is the capture of this plane. Parameters and are alluded to as incomplete relapse coefficients. Parameter speaks to the adjustment in the mean reaction comparing to a unit change in when is held steady. Parameter speaks to the adjustment in the mean reaction comparing to a unit change in when is held consistent. Consider the accompanying case of a different direct relapse model with two indicator variables.

This relapse model is a first request various direct relapse model. This is on the grounds that the most extreme force of the variables in the model is 1. (The relapse plane comparing to this model is appeared in the figure underneath.) Also indicated is a watched information point and the relating arbitrary blunder. The genuine relapse model is normally never known (and thusly the estimations of the arbitrary mistake terms relating to watched information focuses stay obscure). Be that as it may, the relapse model can be evaluated by figuring the parameters of the model for a watched information set. This is clarified in Estimating Regression Models Using Least Squares.

Minimum square system

The most generally perceived system for fitting a relapse line is the procedure for least squares. This framework figures the best-minimizing in order to fit line for the watched data the squares' total of the vertical deviations from each data point to the line (if a point lies on the fitted line unequivocally, then its vertical deviation is 0). Since the deviations are at first squared, then summed, there are no cancelations amidst positive and negative qualities.

Example[9]

The dataset "TVs, Physicians, and Life Expectancy" contains, among diverse variables, the amount of people per TV set and the amount of people per specialist for 40 countries. Since both variables likely mirror the level of riches in every nation, it is sensible to accept that there is some positive relationship between them. Subsequent to evacuating 8 nations with missing qualities from the dataset, the remaining 32 nations have a relationship coefficient of 0.852 for number of individuals per TV set and number of individuals per doctor. The r^2 quality is 0.726 (the connection's square coefficient), showing that 72.6% of the variety in one variable may be clarified by the other.

To see the model's attack to the watched information, one may plot the figured relapse line over the genuine information focuses to assess the outcomes. For this sample, the plot appears to the privilege, with number of people per TV set (the illustrative variable) on the x-hub and number of people per doctor (the ward variable) on the y-hub. While the vast majority of the information focuses are grouped towards the lower left corner of the plot (showing generally couple of people per TV set and per doctor), there are a couple focuses which lie far from the principle group of the information. These focuses are known as outliers, and relying upon their area may have a noteworthy sway on the relapse line.

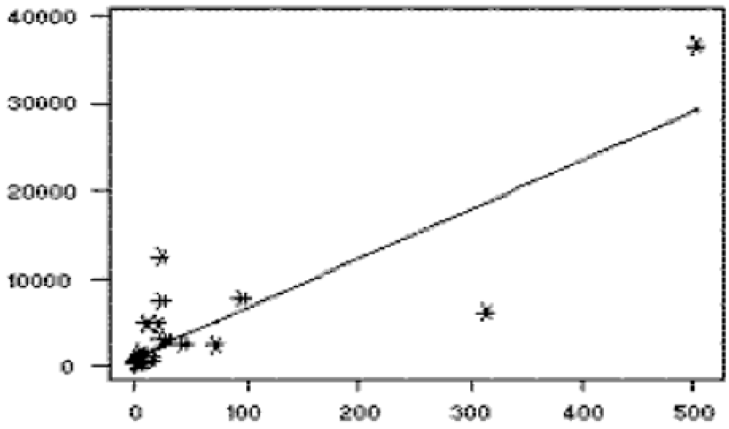

Figure 1: Regression Line formation

Outliers and Influential Observations

After a relapse line has been figured for a gathering of information, a point which lies far from the line (and along these lines has a huge lingering worth) is known as an outlier. Such focuses may speak to incorrect information, or may demonstrate an ineffectively fitting relapse line. On the off chance that a point lies a long way from the other information in the flat course, it is known as an influential perception. The explanation behind this qualification is that these focuses have may have a huge effect on the relapse's incline line. Notice, in the above illustration, the impact of evacuating the perception in the upper right corner of the plot

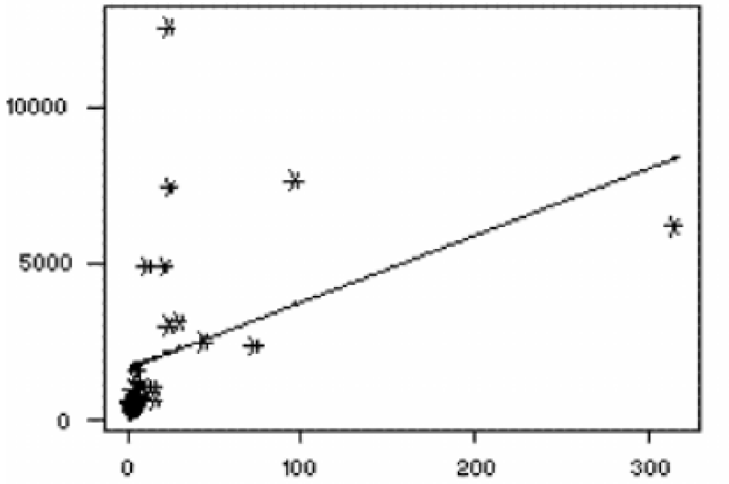

Figure 2: Regression Line

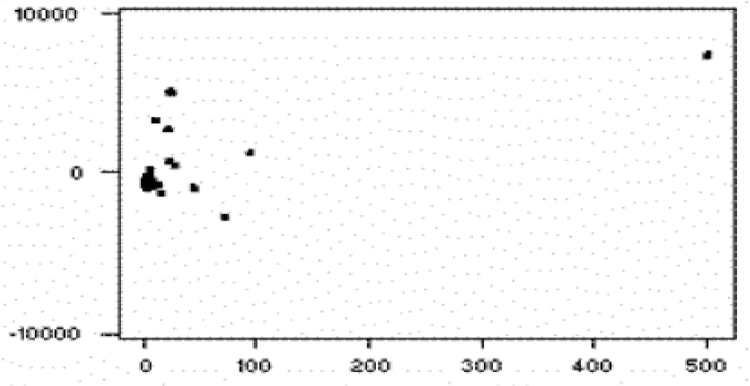

Figure 3: Inlets and outlets

With this powerful perception uprooted, the regression comparison is currently People.Phys = 1650 + 21.3 People.Tel[10]. The connection between the two variables has dropped to 0.427, which decreases the r^2 value to 0.182. With this powerful perception uprooted, less that 20% of the variety in number of individuals per doctor may be clarified by the quantity of individuals per TV. Powerful perceptions are too obvious in the new model, and their effect ought to additionally be researched.

Residuals

When a regression model has been fit to a gathering of information, examination of the residuals (the deviations from the fitted line to the watched qualities) permits the modeller to research the legitimacy of his or her suspicion that a direct relationship exists. Plotting the residuals on the y- pivot against the informative variable on the x-pivot uncovers any conceivable non-straight relationship among the variables, or may alarm the modeller to investigate lurking variables. In our illustration, the remaining plot increases the vicinity of anomalies.

Lurking Variables

In the event that non-direct patterns are noticeable in the relationship between a logical and subordinate variable, there may be other persuasive variables to consider. A lurking variable exists when the relationship between two variables is altogether influenced by the vicinity of a third variable which has not been incorporated into the demonstrating exertion. Since such a variable may be an element of time (for instance, the impact of political or monetary cycles), a time arrangement plot of the information is frequently a helpful apparatus in distinguishing the vicinity of sneaking variables.

RESULTS

At whatever point a direct regression model is fit to a gathering of information, the information's scope should be deliberately watched. Endeavoring to utilize a regression comparison to foresee values outside of this extent is regularly unseemly, and may yield mind boggling answers. This practice is known as extrapolation. Consider, for instance, a direct model which relates weight addition to age for youthful youngsters. Applying such a model to grown-ups, or indeed, even adolescents, would be foolish, since the relationship in the middle of age and weight pick up is not reliable for all age bunches.

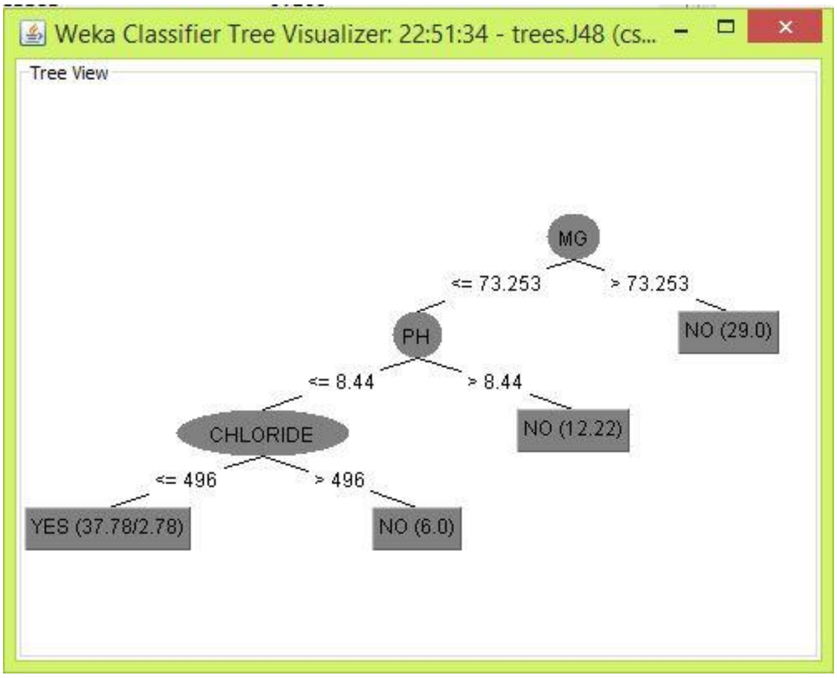

Figure 4: Analysis using J48 tree

CONCLUSION

.

The examination of given data set is done using multiple linear regression a data mining technique. Least square method is used for making the regression line. The elements which fall near to the regression line, which is formed using the standard value are considered good for the drinking purpose. The farther the data fall from the standard value it is considered to be irrelevant for the drinking purpose. And on the basis of this regression method we obtain the water quality.

ACKNOWLEDGEMENT

One of the author (Nishil Shah) would like to thanks Prof. M Venkatesan for the help he has done with our project by providing dataset of the Vellore district water data of wells.

REFERENCES

[1] Analysis of urban water quality based on GIS, Shanshan Li; Huili Gong; Sa Wang; Zhiheng Wang, 2010, Pages: 1 - 4

[2] Application of Comprehensive Water Quality Identification Index in Water Quality Assessment of River, Miao Qun; Gao Ying; Liu Zhiqiang; Tan Xiaohui, 2009, 333 - 337

[3] Notice of Retraction Study on water quality information decision support system, 2010, 197-199

[4] Assessment of spatial-temporal variations in surface water quality of Luzhi river system in Plain River-Net Areas, Suzhou, China, 2011, Pages: 2002-2005

[5] Research on Water Quality Comprehensive Evaluation Index for Water Supply Network Using SOM, Year: 2008, Volume: 2, 6211-624

[6] Demo Abstract: Wind measurements for water quality studies in urban reservoirs, Wan Du; Mo Li; Zikun Xing; Bingsheng He; Chua, L.H.C.; Zhenjiang Li; Yuanqiang Zheng; Pengfei Zhou, Sensing, Communication, and Networking (SECON), 2014 Eleventh Annual IEEE International Conference on, Year: 2014, Pages: 161 - 163, DOI: 10.1109/SAHCN.2014.6990343

[7] Research on water environmental quality evaluation and characteristics analysis of TongHui River, 2011, Volume: 2, Pages: 1066-1069

[8] Discriminant Analysis Method Application in Water Quality Assessment: Take Yinma River as Example, Xin Xin; Lu Wen-xi; Gong Lei, Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on, Year: 2010, Pages: 1 - 3, DOI: 10.1109/ICBBE.2010.5515023

[9] Notice of Retraction Assessment of Surface Water Quality Using Multivariate Statistical Techniques: A Case Study of the Lakes in Wuhan, Yang Yang; Liang Shengwen; Fang Jiande; Zhang Mailang; Zuo Guoxing; Zhang Can; Yang Xue; He Zhen; Hu Xiaojing; Zhong Qiu; Guo Jia; Xiong Li; Liu Deli, Bioinformatics and Biomedical Engineering, (iCBBE) 2011 5th International Conference on, Year: 2011, Pages: 1 - 4, DOI: 10.1109/icbbe.2011.5780731

[10] Identifying Potential Pollution Sources in River Basin via Water Quality Reasoning Based Expert System, Yi Wang; Yuanyuan Wang; Meng Ran; Yu Liu; Zhichao Zhang; Liang Guo; Ying Zhao; Peng Wang, Digital Manufacturing and Automation (ICDMA), 2013 Fourth International Conference on, Year: 2013, Pages: 671 - 674